Name: Itzel Gonzalez

This example is based on the EDA example in Doing Data Science Ch. 2. There are 31 datasets named nyt1.csv, nyt2.csv,…,nyt31.csv, which you can find here: https://github.com/oreillymedia/doing_data_science (https://github.com/oreillymedia/doing_data_science). I have already downloaded the dataset for you to use under the folder: /nytdata

Each file represents one (simulated) day's worth of ads shown and clicks recorded on the New York Times home page in May 2012. Each row represents a single user. There are five columns: age, gender (0=female, 1=male), number impressions, number clicks, and logged-in.

In [5]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Lets start off analyzing one file

In [6]:
```
df = pd.read_csv("dds_ch2_nyt/nyt1.csv") # read-in one file
df.head()
```

Out[6]:

	Age	Gender	Impressions	Clicks	Signed_In
0	36	0	3	0	1
1	73	1	3	0	1
2	30	0	3	0	1
3	49	1	3	0	1
4	47	1	11	0	1

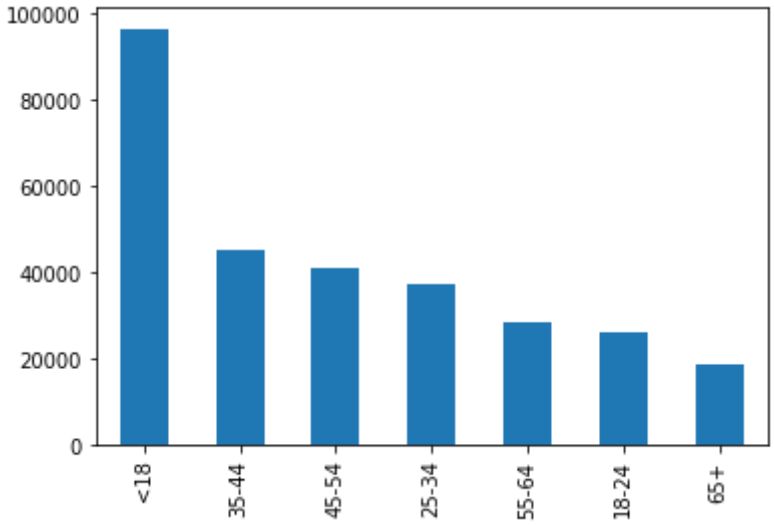
Once you have the data loaded, it's time for some EDA:

Create a new variable, age_group, that categorizes users as "<18", "18-24", "25-34", "35-44", "45-54", "55-64", and "65+".

In [7]:
```
bins= [0,18,25,35,45,55,65,108]
labels = ['<18','18-24','25-34','35-44','45-54','55-64','65+']
df['age_group'] = pd.cut(df['Age'], bins=bins, labels=labels, right=Fals
e)
```

In [8]:
```
df['age_group'].value_counts().plot(kind='bar')
```

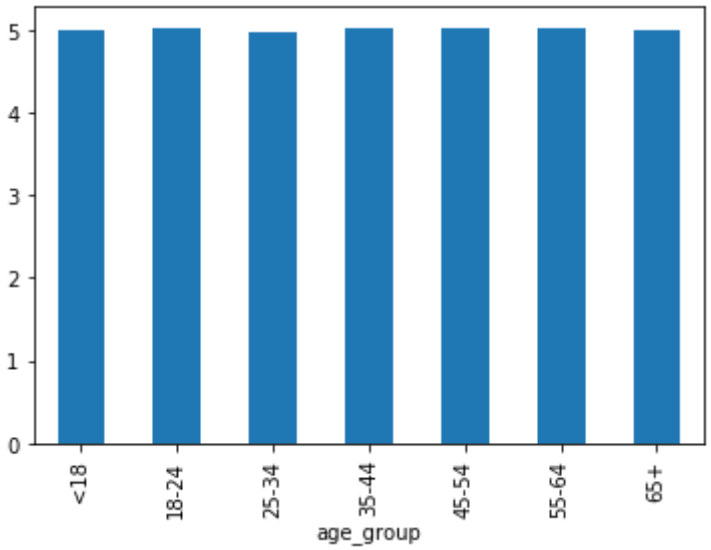
Out[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8f2ae53ba8>
```



Plot the distributions of number impressions and click-through-rate for these six age categories.

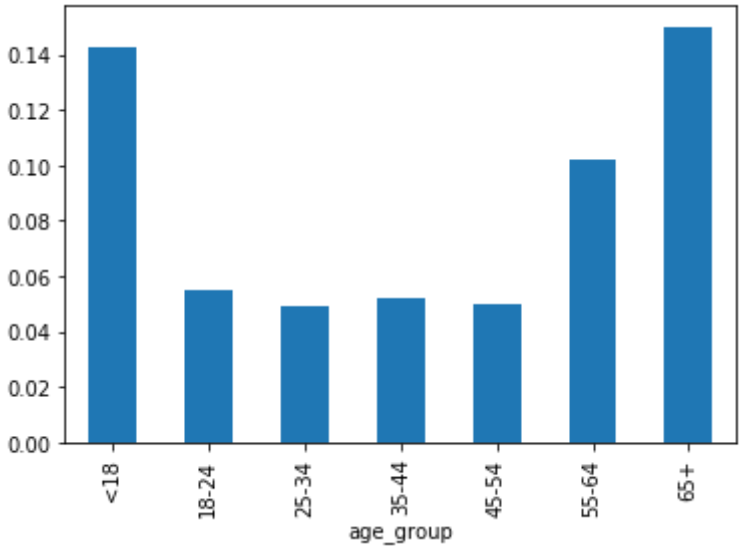
In [12]:
```
df.groupby("age_group").mean()["Impressions"].plot(kind='bar')
```

Out[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8f2a5ef588>
```



In [13]:
```
df.groupby("age_group").mean()["Clicks"].plot(kind='bar')
```

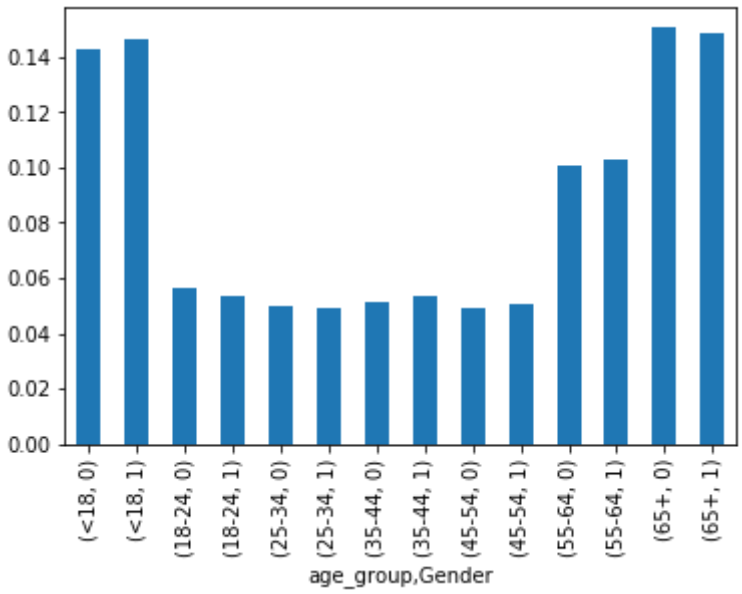
Out[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8f2a4473c8>
```



Explore the data and make visual and quantitative comparisons across user segments/demographics (<18-year-old males versus < 18-year-old females or logged-in versus not, for example).

In [14]:
```
# TODO
df.groupby(["age_group", "Gender"]).mean()["Clicks"].plot(kind='bar')
```

Out[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8f2a43aeb8>
```



We analyzed just one file, but the dataset includes 31 files (

In [15]:
```
import glob # used to read mutliple-files
files = glob.glob('dds_ch2_nyt/nyt*.csv')
dfs = []
for file in files:
    df = pd.read_csv(file)
    df['filename'] = file
    dfs.append(df)
df = pd.concat(dfs, ignore_index=True)
df
```

Out[15]:

	Age	Gender	Impressions	Clicks	Signed_In	filename
0	63	1.0	2.0	0.0	1.0	dds_ch2_nyt/nyt17.csv
1	0	0.0	7.0	0.0	0.0	dds_ch2_nyt/nyt17.csv
2	0	0.0	8.0	0.0	0.0	dds_ch2_nyt/nyt17.csv
3	0	0.0	4.0	0.0	0.0	dds_ch2_nyt/nyt17.csv
4	61	1.0	6.0	0.0	1.0	dds_ch2_nyt/nyt17.csv
...	...	...	...	...	...	...
8363192	72	0.0	3.0	0.0	1.0	dds_ch2_nyt/nyt11.csv
8363193	0	0.0	2.0	0.0	0.0	dds_ch2_nyt/nyt11.csv
8363194	0	0.0	5.0	0.0	0.0	dds_ch2_nyt/nyt11.csv
8363195	22	1.0	5.0	1.0	1.0	dds_ch2_nyt/nyt11.csv
8363196	17	1.0	NaN	NaN	NaN	dds_ch2_nyt/nyt11.csv

8363197 rows × 6 columns

Analyze trends over time since we now have a historic view of the data over 31 days.

In [17]:
```
df["Date"]=df["filename"].replace(["dds_ch2_nyt/nyt1.csv",
"dds_ch2_nyt/nyt2.csv","dds_ch2_nyt/nyt3.csv",
"dds_ch2_nyt/nyt4.csv","dds_ch2_nyt/nyt5.csv",
"dds_ch2_nyt/nyt6.csv","dds_ch2_nyt/nyt7.csv",
"dds_ch2_nyt/nyt8.csv","dds_ch2_nyt/nyt9.csv",
"dds_ch2_nyt/nyt10.csv","dds_ch2_nyt/nyt11.csv",
"dds_ch2_nyt/nyt12.csv","dds_ch2_nyt/nyt13.csv",
"dds_ch2_nyt/nyt14.csv","dds_ch2_nyt/nyt15.csv",
"dds_ch2_nyt/nyt16.csv","dds_ch2_nyt/nyt17.csv",
"dds_ch2_nyt/nyt18.csv","dds_ch2_nyt/nyt19.csv",
"dds_ch2_nyt/nyt20.csv","dds_ch2_nyt/nyt21.csv",
"dds_ch2_nyt/nyt22.csv","dds_ch2_nyt/nyt23.csv",
"dds_ch2_nyt/nyt24.csv","dds_ch2_nyt/nyt25.csv",
"dds_ch2_nyt/nyt26.csv","dds_ch2_nyt/nyt27.csv",
"dds_ch2_nyt/nyt28.csv","dds_ch2_nyt/nyt29.csv",
"dds_ch2_nyt/nyt30.csv","dds_ch2_nyt/nyt31.csv"],

["1","2","3","4","5","6","7","8","9","10","11","12",
 "13","14","15","16","17","18","19","20","21","22",
 "23","24","25","26","27","28","29","30","31"])

df
```

Out[17]:

	Age	Gender	Impressions	Clicks	Signed_In	filename	Date
0	63	1.0	2.0	0.0	1.0	dds_ch2_nyt/nyt17.csv	17
1	0	0.0	7.0	0.0	0.0	dds_ch2_nyt/nyt17.csv	17
2	0	0.0	8.0	0.0	0.0	dds_ch2_nyt/nyt17.csv	17
3	0	0.0	4.0	0.0	0.0	dds_ch2_nyt/nyt17.csv	17
4	61	1.0	6.0	0.0	1.0	dds_ch2_nyt/nyt17.csv	17
...	...	...	...	...	...	...	...
8363192	72	0.0	3.0	0.0	1.0	dds_ch2_nyt/nyt11.csv	11
8363193	0	0.0	2.0	0.0	0.0	dds_ch2_nyt/nyt11.csv	11
8363194	0	0.0	5.0	0.0	0.0	dds_ch2_nyt/nyt11.csv	11
8363195	22	1.0	5.0	1.0	1.0	dds_ch2_nyt/nyt11.csv	11
8363196	17	1.0	NaN	NaN	NaN	dds_ch2_nyt/nyt11.csv	11

8363197 rows × 7 columns

In [20]:
```
df_by_date = df.set_index("Date")
df_by_date
```

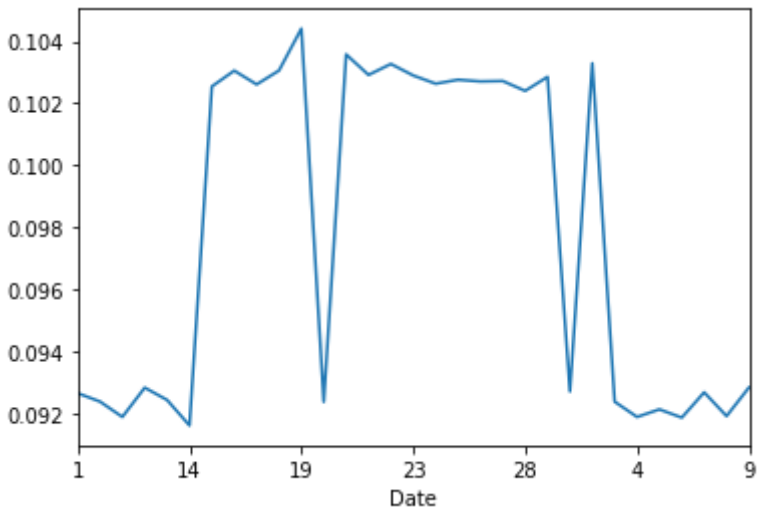
Out[20]:

Date	Age	Gender	Impressions	Clicks	Signed_In	filename
17	63	1.0	2.0	0.0	1.0	dds_ch2_nyt/nyt17.csv
17	0	0.0	7.0	0.0	0.0	dds_ch2_nyt/nyt17.csv
17	0	0.0	8.0	0.0	0.0	dds_ch2_nyt/nyt17.csv
17	0	0.0	4.0	0.0	0.0	dds_ch2_nyt/nyt17.csv
17	61	1.0	6.0	0.0	1.0	dds_ch2_nyt/nyt17.csv
...	...	...	...	...	...	...
11	72	0.0	3.0	0.0	1.0	dds_ch2_nyt/nyt11.csv
11	0	0.0	2.0	0.0	0.0	dds_ch2_nyt/nyt11.csv
11	0	0.0	5.0	0.0	0.0	dds_ch2_nyt/nyt11.csv
11	22	1.0	5.0	1.0	1.0	dds_ch2_nyt/nyt11.csv
11	17	1.0	NaN	NaN	NaN	dds_ch2_nyt/nyt11.csv

8363197 rows × 6 columns

In [21]:
```
df_by_date.groupby(["Date"]).mean()["Clicks"].plot()
```

Out[21]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8f2a366080>
```



In [ ]:
```

```